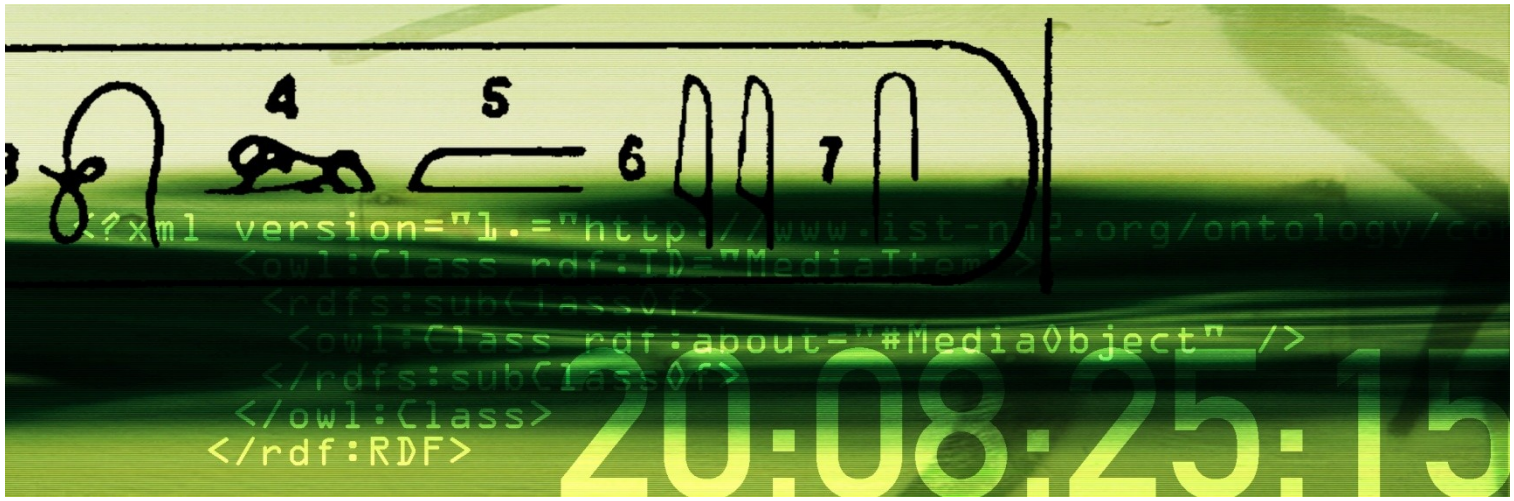
JOANNEUM RESEARCH

Institute of Information Systems & Information Management



Towards Opinion Mining Through Tracing Discussions on the Web

Social Data on the Web Workshop (SDoW 2008)

Michael Hausenblas, Selver Softic

2008-10-27

www.joanneum.at

JOANNEUM RESEARCH

Agenda

- Motivation
- Background
- Representing Discussions and Opinions
- Discussion Tracing
- Data Acquisition
- Analyser
- Preliminary Results
- Future Work

JOANNEUM RESEARCH

Motivation

"*Current search technology is unable to satisfy any complex queries requiring information integration such as analysis, prediction, scheduling, etc. An example of such* ***integration-based tasks is opinion mining regarding products or services****. While there have been some successes in opinion mining with pure sentiment analysis, it is often the case that* ***users like to know what specific aspects of a product or service are being described in positive or negative terms*** *and to have the search results appear aggregated and organized.*"

Peter Mika, Yahoo! Research [1]

JOANNEUM RESEARCH

Motivation

- Products or services are often discussed by customers on the Web
- Online communities are interesting for market analysis and research
- Official company sites usually tell a certain side of the story
- Valuable data relevant for market research on the Web is neither easy accessible nor processable
- Time expenses to collect and evaluate data needed for a better market understanding are still tremendous

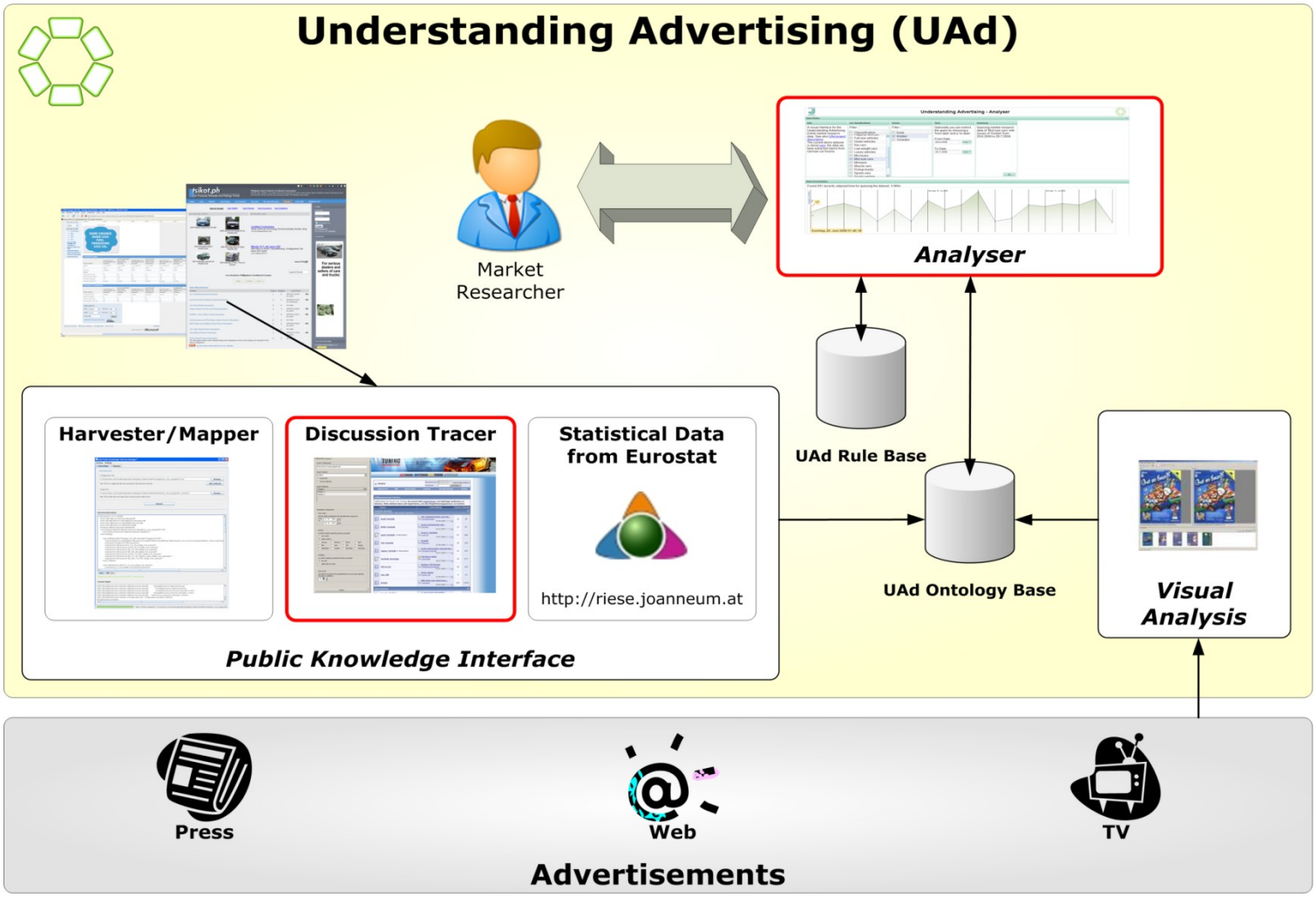
JOANNEUM RESEARCH

Background

- Understanding Advertising (UAd) project aiming at developing a methodology allowing a market researcher to understand a certain market.
- Twofold analysis:
  - by visual interpretation of advertisements (from print media, Web and TV)
  - by using information available on the Web (so called public knowledge interface, PKI)
- We are after: **tracing and understanding discussions on the Web**

JOANNEUM RESEARCH

Background



www.joanneum.at

JOANNEUM RESEARCH

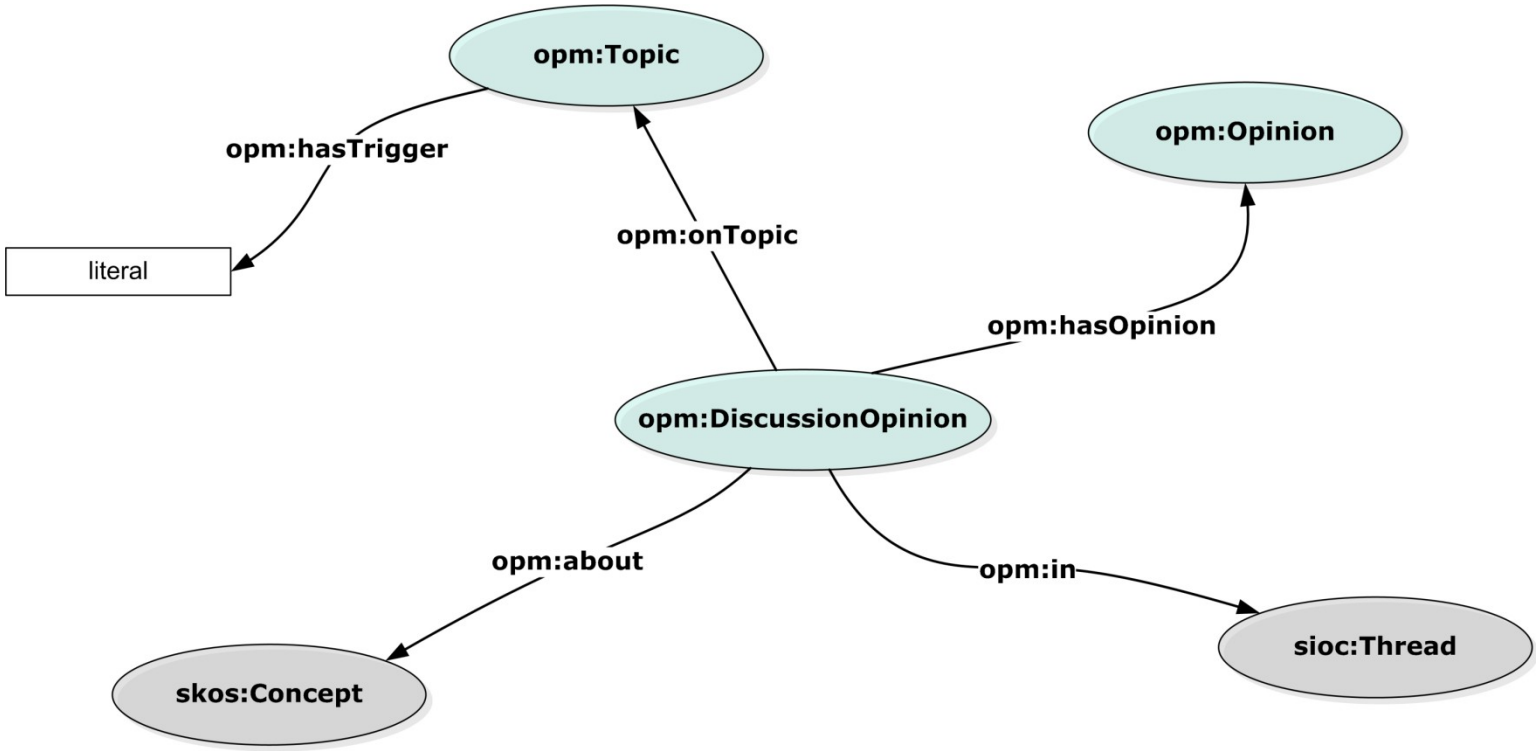
Background

- Sentiment analysis, simple “pro” and “cons” classification for documents
- Often used so far:
  - Natural Language Processing (NLP)
  - machine learning techniques
- The opinion mining workflow usually comprises three major phases:
  - extraction
  - structuring
  - summarisation

Representing Discussions and Opinions

- Assumptions/Goals
  - Represent the discussions in a machine-interpretable way and enhance it with domain semantics
  - Modeling of the opinions in a discussion should be compliant to the Web of Data
  - Reuse of an existing vocabularies
  - Orienting the opinion holder context on domain semantics by exploiting linked datasets (such as DBpedia)

Representing Discussions and Opinions



opm: <http://sw.joanneum.at/uad/u-opm/schema/core-u-opm.rdf#>
sioc: <http://rdfs.org/sioc/ns#>
skos: <http://www.w3.org/2004/02/skos/core#>

Representing Discussions and Opinions

@prefix **sioc**: <http://rdfs.org/sioc/ns#> .

@prefix **dcterms**: <http://purl.org/dc/terms/#> .

@prefix **foaf**: <http://xmlns.com/foaf/0.1/> .

@prefix **rdf**: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

@prefix **rdfs**: <http://www.w3.org/2000/01/rdf-schema#> .

@prefix **xsd**: <http://www.w3.org/2001/XMLSchema#> . <http://www.automotiveforums.com/vbulletin/showpost.php?s=1439f714b241984e4daad2eb2450f1ff&p=1903642&postcount=3> a **sioc:Post;**

dcterms:created "2004-07-06";

dcterms:title "Dead Battery - Can't Get Into Trunk!!!";

sioc:content

"Use a battery charger that charges the car battery thru the cigarette lighter. There are a lot of Volkswagen OEM…."

sioc:has_container <http://www.automotiveforums.com/vbulletin/showthread.php?s=1439f714b241984e4daad2eb2450f1ff&t=238034>;

sioc:has_creator <http://www.automotiveforums.com/vbulletin/member.php?s=1439f714b241984e4daad2eb2450f1ff&u=192327>;

sioc:has_reply <http://www.automotiveforums.com/vbulletin/showpost.php?s=1439f714b241984e4daad2eb2450f1ff&p=2927187&postcount=4>, <http://www.automotiveforums.com/vbulletin/showpost.php?s=1439f714b241984e4daad2eb2450f1ff&p=3289293&postcount=5> .

Representing Discussions and Opinions

```
@prefix : <http://sw.joanneum.at/uad/cars/topics#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix opm: <http://sw.joanneum.at/uad/u-opm/schema/core-u-opm.rdf#>.
<http://sw.joanneum.at/uad/cars/opinions#do1>
    a opm:DiscussionOpinion;
    opm:in <http://www.automotiveforums.com/vbulletin/showthread.php?s=1439f714b241984e4daad2eb2450f1ff&t=591194>;
    opm:onTopic :performance_and_problems,
    opm:about <http://dbpedia.org/resource/Jaguar_S-Type> .
```

Representing Discussions and Opinions

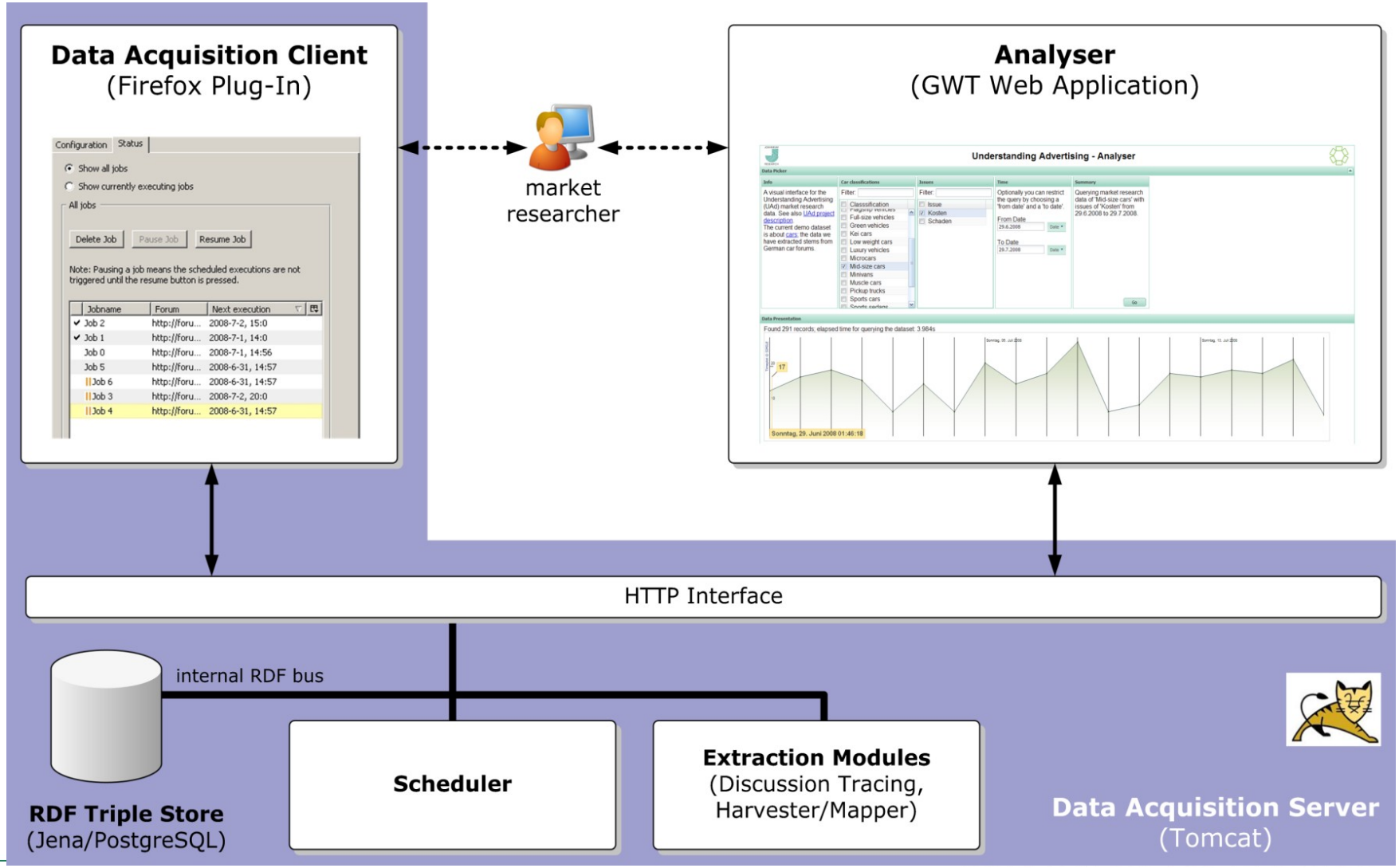
```
:performance_and_problems
      a opm:Topic;
        dc:subject "performance and problems";
        opm:hasTrigger
            „key,
            "alarm,
            „batter“,
            "transmission",
            "trouble",
            "trunk,
             …
            "window" .
```

JOANNEUM RESEARCH

Discussion Tracing

- Data Acquisition module
  - Includes harvesting, RDFising, interlinking and opinion generation
- UAd Analyser
  - allows to query and access the data.

Discussion Tracing

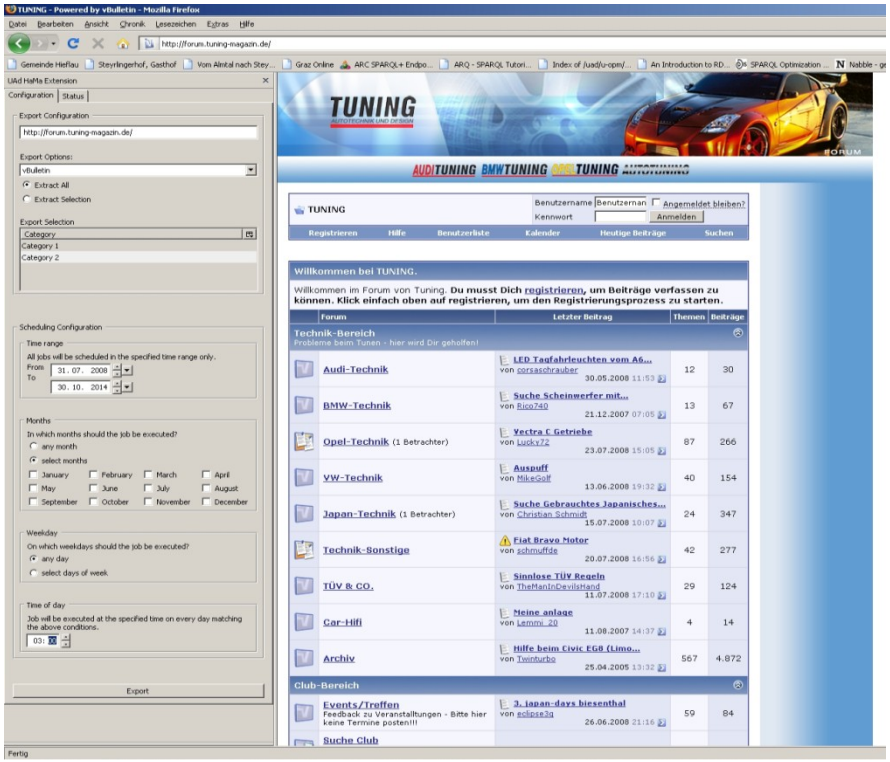


JOANNEUM RESEARCH

Data Acqusition

Performed in three phases:

1. RDFising data from Web-based discussion forums using SIOC
2. Interlinking with domain concepts (DBPedia)
3. Opinion Generation



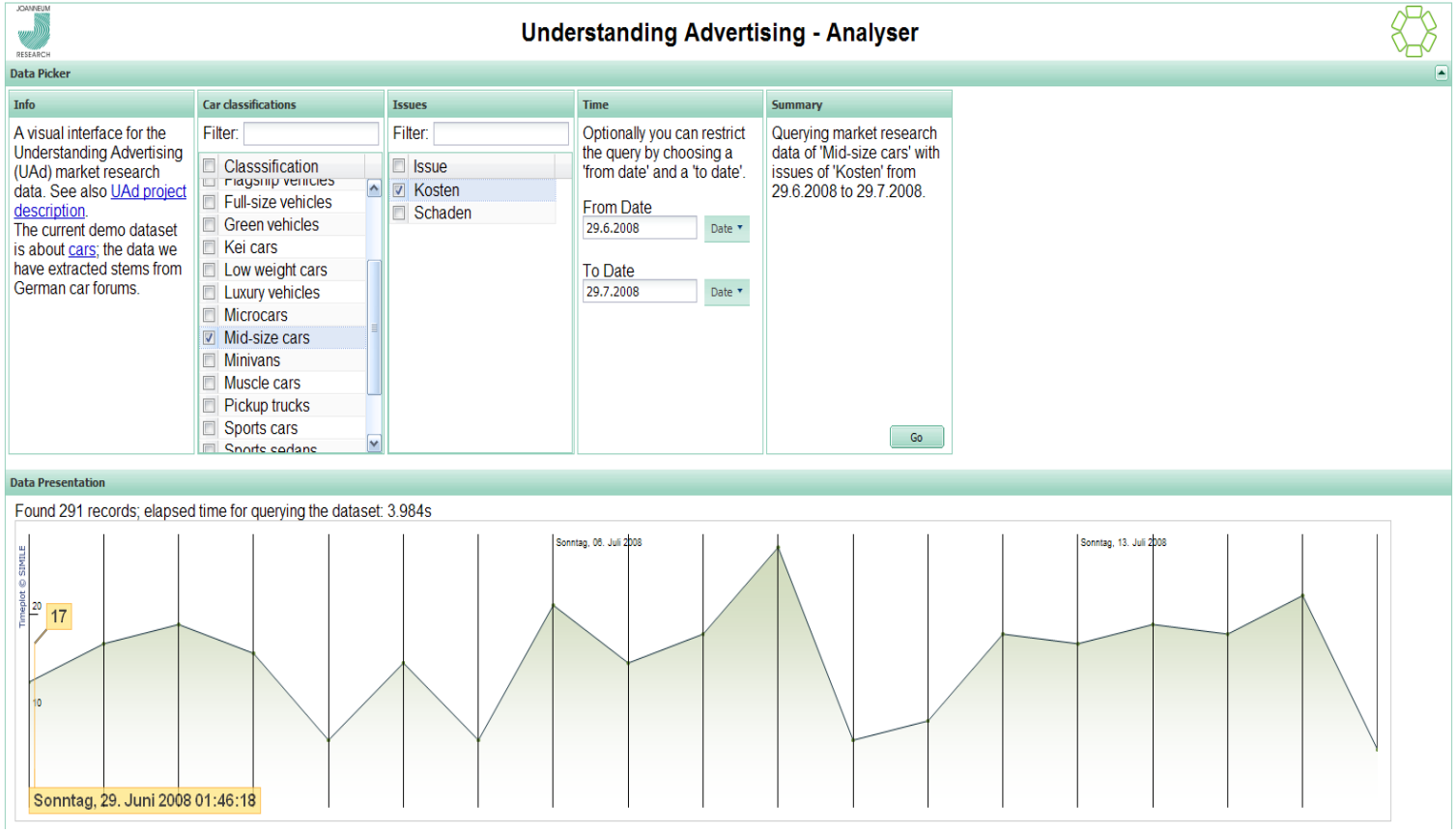
JOANNEUM RESEARCH

Data Acqusition

- Opinion Mining Core Ontology serves as nexus
- Extracted SIOC data is reusable
- Client/Server
  - FireFox Add-On (UAd Acqusition Client)
  - Tomcat + PostgresSQL + Servlets (UAd Acqusition Server)
- Automated extraction jobs
- DBpedia choice is mandatory
- Extraction tasks
  - One configuration file per forum type
- Interlinking occurs manually at the moment

JOANNEUM RESEARCH

Analyser



http://uad.joanneum.at/analyser/

www.joanneum.at

JOANNEUM RESEARCH

Analyser

- Allows market researcher examination of data gathered by the acquisition task for a certain domain
- Reflects information
  - Post count respectively a date
  - Author diversity on a certain day
  - Posts with references for detailed browsing

```
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix utop: <http://sw.joanneum.at/uad/cars/topics#> .
prefix opm: <http://sw.joanneum.at/uad/u-opm/schema/core-u-opm.rdf#>

SELECT * FROM <http://sw.joanneum.at/uad>
WHERE {
 ?do a opm:DiscussionOpinion ;
      opm:about ?about;
      opm:in ?in ;
      opm:onTopic utop:performance_and_problems .
 ?about owl:sameAs <http://dbpedia.org/resource/Alfa_Romeo_156> .
}
```

JOANNEUM RESEARCH

Preliminary Results

- Baseline comparion with Lucene using precision and recall
- Reference data set:
  - 1000 Posts
  - Dbpedia domain (cars)
- Working data set:
  - 60 posts
  - 20 per car type
  - 2 categories („*popularity*“ and „*performance and problems*“)
  - Two out of three car types belong to compatible category from DBPedia
- Simple and extended queries

JOANNEUM RESEARCH

Preliminary Results

		Lucene		UAd Analyser	
		"performance and problems"	*"popularity"*	*"performance and problems"*	*"popularity"*
Precision	simple	0.4	1	0.76	0.86
	extended	0.2–0.62	0.56–0.86		
Recall	simple	0.1	0.05	0.95	0.6
	extended	0.05–0.8	0.3–0.7		

JOANNEUM RESEARCH

Future Work

- Enhancing sentiment classification and identification
  - NLP
  - SVM
  - SentiWordNet
- Automation of interlinking
  - Equivalence mining
- Using URIs as triggers

JOANNEUM RESEARCH

Questions?

THANK YOU!

www.joanneum.at

JOANNEUM RESEARCH

References

1. P. Mika. Microsearch: An Interface for Semantic Search. In Proc. of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008) , Tenerife, Spain, volume 334 of CEUR Workshop Proceedings. CEUR-WS.org, 2008.
2. W. Halb, Y. Raimond, andM. Hausenblas. Building Linked Data For Both Humans and Machines. In WWW 2008 Workshop: Linked Data on the Web (LDOW2008), Beijing, China, 2008.
3. M. Hausenblas, W. Halb, and Y. Raimond. Scripting User Contributed Interlinking. In 4th Workshop on Scripting for the Semantic Web (SFSW08), Tenerife, Spain, 2008.
4. S. Kim and E. Hovy. Automatic identification of pro and con reasons in online reviews. In Proceedings of the COLING/ACL on Main conference poster sessions, pages 483–490, Morristown, NJ, USA, 2006. Association for Computational Linguistics.
5. M. Hu and B. Liu. Mining opinion features in customer reviews. In American Association for Artificial Intelligence at AAAI-04, 2004.
6. M. Hu and B. Liu. Mining and summarizing customer reviews. In Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining at KDD-2004, pages 168–177, 2004.
7. K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In WWW2003 - The Twelfth International World Wide Web Conference, Budapest, HUNGARY, 2003.
8. N. Kobayashi, K. Inui, and Y. Matsumoto. Opinion Mining from Web Documents: Extraction and Structurization. Informational and Media Technologies 2(1), 12(1):326–337, 2007.
9. A. Ghose, P. Ipeirotis, and A. Sundararajan. Opinion Mining using Econometrics: A Case Study on Reputation Systems. In Proceedings of the Association for Computational Linguistics (ACL), 2007.
10. M. Gamon and A. Aue. Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms. In Proceedings of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing, 2005.

JOANNEUM RESEARCH

References

11. M. Hausenblas and H. Rehatschek. mle: Enhancing the Exploration of Mailing List Archives Through Making Semantics Explicit. In Semantic Web Challenge 2007 at the 6th International Semantic Web Conference (ISWC07), Busan, South Korea, 2007.
12. S. Fernandez, D. Berrueta, and J.E. Labra. Mailing Lists Meet The Semantic Web. In Proc. of the BIS 2007 Workshop on Social Aspects of the Web, Poznan, Poland, 2007.
13. S. Fernandez, F. Giasson, and K. Idehen. SIOC Ontology: Applications and Implementation Status. http://www.sioc-project.org/applications#creating-api, 2007.
14. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A Nucleus for a Web of Open Data. In The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, pages 722–735, 2007.
15. J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. Inf. Process. Manage., 43(4):866–886, 2007.
16. A. Esuli. Opinion Mining. Presentation slides, Language and Intelligence Reading Group, June 14, 2006, Pisa, Italy, Istituto di Scienza e Tecnologie dell' Informazione Consiglio Nazionale delle Ricerche, 2006.
17. B. Liu. Opinion Mining and Summarization, Sentiment Analysis. Presentation slides, Tutorial given at WWW-2008, April 21, 2008 in Beijing, China, Department of Computer Science University of Illinois at Chicago, 2008.
18. T. Berners-Lee. Linked Data. http://www.w3.org/DesignIssues/LinkedData.html, 2007.
19. C. Bizer, T. Heath, D. Ayers, and Y. Raimond. Interlinking Open Data on the Web (Poster). In 4th European Semantic Web Conference (ESWC2007), pages 802–815, 2007.

JOANNEUM RESEARCH

References

20. T. Heath and E. Motta. Revyu.com: a Reviewing and Rating Site for the Web of Data. In The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, pages 895–902, 2007.
21. Semantic Web Deployment Working Group. SKOS Simple Knowledge Organization System Reference. W3C Working Draft, Semantic Web Deployment Working Group, 2008.
22. G. K. Zipf. Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology. Addison-Wesley, 1949.
23. A. Esuli and F. Sebastiani. SentiWordnet: A Publicly Available Lexical Resource for Opinion Mining. In 5th Conference on Language Resources and Evaluation (May 22–28, 2006), Genova, Italy, 2006.